

# Stratification of patient trajectories using covariate latent variable models

Kieran R. Campbell & Christopher Yau  
Wellcome Trust Centre for Human Genetics  
University of Oxford  
kieran.campbell@sjc.ox.ac.uk

## Abstract

Standard models assign disease progression to discrete categories or *stages* based on well-characterized clinical markers. However, such a system is potentially at odds with our understanding of the underlying biology, which in highly complex systems may support a (near-)continuous evolution of disease from inception to terminal state. To learn such a continuous disease score one could infer a latent variable from dynamic “omics” data such as RNA-seq that correlates with an outcome of interest such as survival time. However, such analyses may be confounded by additional data such as clinical covariates measured in electronic health records (EHRs). As a solution to this we introduce *covariate latent variable models*, a novel type of latent variable model that learns a low-dimensional data representation in the presence of two (asymmetric) views of the same data source. We apply our model to TCGA colorectal cancer RNA-seq data and demonstrate how incorporating microsatellite-instability (MSI) and metastatic status as external covariates allows us to identify genes that stratify patients on an immune-response trajectory. Finally, we propose an extension termed *Covariate Gaussian Process Latent Variable Models* for learning nonparametric, nonlinear representations.

## 1 Introduction

There exists a set of physical processes with an assumed temporal component but where precise measurement of times associated with events is precluded or impossible. Such ideas have recently flourished in the field of single-cell genomics, where cells will undergo some dynamic process such as differentiation but in which the destructive measurement of gene expression precludes physical measurement of the progression itself. Consequently, the progression is artificially inferred from the measured expression data as a *pseudotime* (e.g. [1, 2]), which in a statistical sense is akin to inference of a one-dimensional latent variable model.

This situation also arises in the case of disease staging and survival analysis such as when a patient presents to a clinic with a disease of unknown progression. Typically, the patient will be assigned a discrete disease *stage* after possibly invasive tests and/or surgery. The discrete nature of such staging is at odds with accepted knowledge of the underlying biology, which is consistent with a more continuous evolution of disease progression such as gradual changes in gene expression. Furthermore, such evolution is confounded by underlying population heterogeneity, where the evolution of molecular features along the trajectory may differ depending on external patient phenotypes, such as age and sex or molecular phenotypes such as mutations of a particular gene (Figure 1A).

As a proof-of-concept solution to such issues we propose *Covariate Latent Variable Models* (C-LVMs), a novel type of latent variable model similar to factor analysis in which the evolution of various dynamic genomic observables (such as gene expression) is allowed to vary according to a secondary set of covariates. Such a model represents latent variable inference from two *views* of the

same sample but where the relationship between each view and the latent variables is asymmetric. Formulated as a Bayesian hierarchical model we are further able to extract interactions between the patient trajectory and covariates, simultaneously providing insight into the underlying biology. We apply our model to lipid metabolism gene expression for the TCGA colorectal cancer dataset using microsatellite instability and metastatic status as covariates and extract a trajectory consistent with known markers of colorectal cancer prognosis. Finally, we propose a nonlinear, nonparametric extension and discuss the relationship to Gaussian Process Latent Variable Models.

## 2 Methods

We begin with an  $N \times G$  data matrix  $\mathbf{Y}$  where  $y_{ig}$  denotes the  $i^{th}$  entry in the  $g^{th}$  column for  $i \in 1, \dots, N$  samples and  $g \in 1, \dots, G$  features. Such a matrix would correspond to the measurement of a dynamic molecular process that we might reasonably expect to show continuous evolution as a disease progresses such as gene expression corresponding to a particular pathway. It is then trivial to learn a one-dimensional linear embedding that would be our “best guess” of such progression via a factor analysis model:

$$y_{ig} = c_g t_i + \epsilon_{ig}, \quad \epsilon_{ig} \sim N(0, \tau_g^{-1}) \quad (1)$$

where  $t_i$  is the latent measure of progression for sample  $i$  and  $c_g$  is the factor loading for feature  $g$  which essentially describes the evolution of  $g$  along the patient trajectory.

However, it is conceivable that the evolution of feature  $g$  along the trajectory is not identical for all samples but is instead affected by a set of external covariates. Such covariates may correspond to patient phenotypes such as age or sex, EHR entries such as blood pressure or additional molecular data such as the mutation status of a particular gene. Note that we expect such features to be “static” and not necessarily correlate with the trajectory itself.

Introducing the  $N \times P$  covariate matrix  $\mathbf{X}$  with the entry in the  $i^{th}$  row and  $p^{th}$  column given by  $x_{ip}$ , we allow such measurements to perturb the factor loading matrix

$$c_g \rightarrow c_g + \sum_{p=1}^P \beta_{pg} x_{ip} \quad (2)$$

where  $\beta_{pg}$  quantifies the effect of covariate  $p$  on the evolution of feature  $g$ . Despite  $\mathbf{Y}$  being column-centred we need to reintroduce gene and covariate specific intercepts to satisfy the model assumptions, giving a likelihood of

$$y_{ig} = \eta_g + \sum_{p=1}^P \alpha_{pg} x_{ip} + \left( c_g + \sum_{p=1}^P \beta_{pg} x_{ip} \right) t_i + \epsilon_{ig}, \quad \epsilon_{ig} \sim N(0, \tau_g^{-1}) \quad (3)$$

Our goal is inference of  $t_i$  that is useful for patient stratification along with  $\beta_{pg}$  which is informative of novel interactions between continuous trajectories and external covariates. Consequently we place a sparse Bayesian prior on  $\beta_{pg}$  of the form  $\beta_{pg} \sim N(0, \chi_{pg}^{-1})$  where the posterior of  $\chi_{pg}$  is informative of the model’s belief that  $\beta_{pg}$  is non-zero.

To understand this model it helps to consider the distribution of  $\mathbf{Y}$  marginalised over the mapping  $\{c_g, \alpha_{pg}, \beta_{pg}\} \forall p, g$  with priors  $c_g \sim N(0, \tau_c^{-1})$  and  $\alpha_{pg} \sim N(0, \tau_\alpha^{-1})$ , then if  $\mathbf{y}_g$  denotes the column vectors of  $\mathbf{Y}$  and similarly  $\mathbf{x}_p$  for  $\mathbf{X}$ ,  $[\mathbf{t}]_i = t_i$ ,  $\mathbf{1}_N$  is the column vector of ones and  $\odot$  denotes the element-wise product, then

$$p(\mathbf{y}_g | \mathbf{X}, \mathbf{t}, \eta_g, \tau_g, \tau_c, \tau_\alpha, \chi_{pg}) \sim N\left(\eta_g \mathbf{1}_N, \Sigma^{(g)}\right) \quad (4)$$

where

$$\Sigma^{(g)} = \tau_g^{-1} \mathbf{1}_N \mathbf{1}_N^T + \tau_\alpha^{-1} \mathbf{X} \mathbf{X}^T + \tau_c^{-1} \mathbf{t} \mathbf{t}^T + \sum_p \chi_{pg}^{-1} (\mathbf{x}_p \odot \mathbf{t})(\mathbf{x}_p \odot \mathbf{t})^T. \quad (5)$$

We therefore see that the addition of the covariates adds extra terms to the covariance matrix corresponding to *perturbations* of the latent variables with the covariates. Consequently, the scale

on which  $\mathbf{x}_p$  is defined needs carefully calibrated. Furthermore, it is possible to extend the latent variable matrix to have dimension larger than 1 giving a novel dimensionality reduction technique for visualisation, though additional rotation issues arise.

### 3 Results

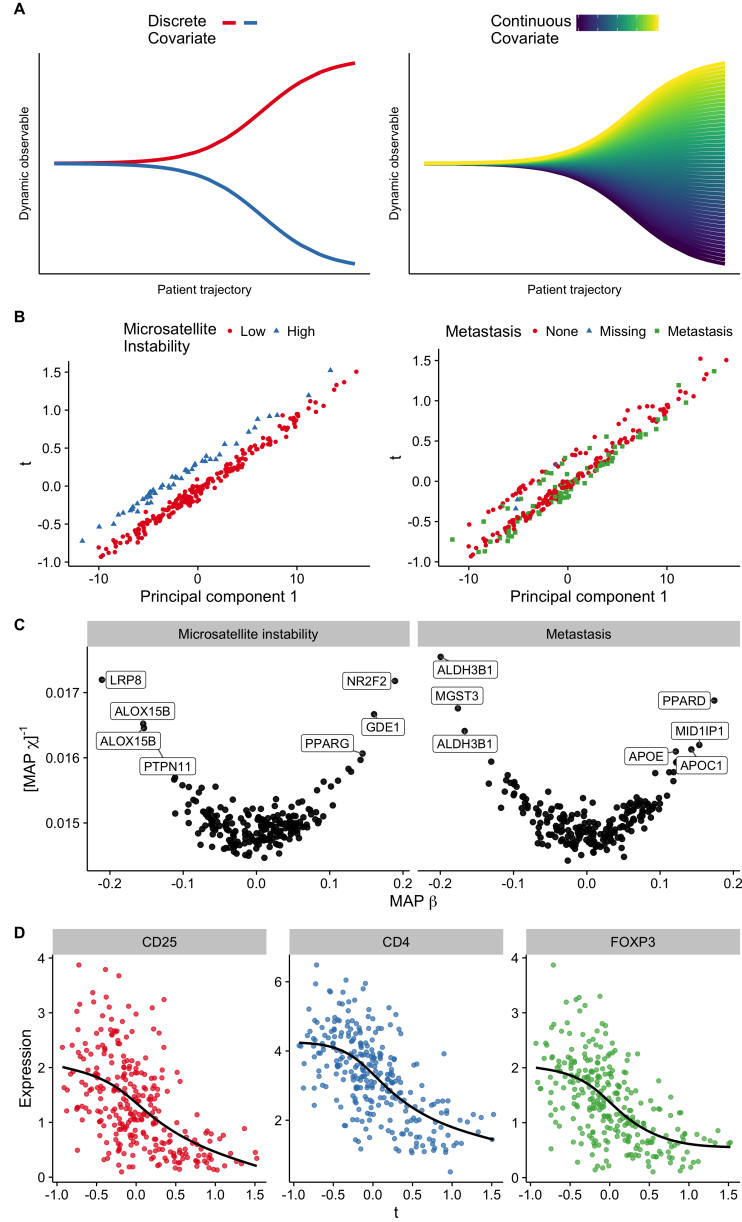


Figure 1: Covariate latent variable models applied to lipid metabolism genes in the TCGA colorectal cancer RNAseq dataset using microsatellite instability and metastatic status as covariates. **A** C-LVMs can be applied to infer trajectories in the presence of discrete covariates (left) or continuous covariates (right). **B** Comparison of inferred  $t$  to principal component 1. Microsatellite instability status results in a small but significant shift in the latent variables. **C** Posterior variance  $\chi^{-1}$  compared to the posterior coefficient values  $\beta$  identify lipid metabolism genes that interact with the covariates along the trajectory. **D** Expression plots of T cell regulator markers (*CD25*, *CD4* and *FOXP3*) show regulation along the trajectory. T cell regulator status is known to be correlated with prognosis.

We applied our method to a recent transcript-level quantification [3] of the TCGA colorectal cancer dataset [4]. We used gene expression for genes annotated in the lipid metabolism pathway as  $\mathbf{Y}$  and whether the tumour showed high microsatellite instability (MSI) or had metastasised as the covariates for  $\mathbf{X}$ . As tumours grow they face increased metabolic demands [5], leading us to believe such a trajectory would leave a footprint in genes associated with metabolism, while tumours with MSI are known to give rise to distinct phenotypes [6].

The results may be seen in Figures 1B-D. The strong sparsity priors on  $\beta$  suggest the effect of the covariates  $\mathbf{X}$  on the inferred  $\mathbf{t}$  will be minimal and should be similar to traditional factor analysis models; however, Figure 1B clearly shows MSI status having a discernible shift on the inferred latent variables. While this effect is small in the dataset examined it is conceivable that larger confounding effects may exist in other datasets. We subsequently examined the posterior values of  $\chi^{-1}$  and  $\beta$  for each coefficient and gene to discover any interactions between the lipid-metabolic trajectory and the covariates included (Figure 1C). This identified genes such as *NRF2*, *ALDH3B1* and *PPARD* all associated with colorectal cancer outcomes [7–9].

Finally, we sought to calibrate our inferred trajectory with some external measure of progression. Survival analysis in TCGA data is problematic - measurements of survival are taken from initial prognosis with scarce recording of the assay timing relative to this. Furthermore, in the colorectal cancer dataset used 240 / 283 (85%) patients have no survival information recorded at all. Consequently, we sought to compare our trajectory with a genomic measure of prognosis, namely *FOXP3+* regulatory T-cell (Treg) status which is associated with poor colorectal cancer prognosis [10]. We examined the expression of three Treg markers along  $\mathbf{t}$  (*CD25*, *CD4* and *FOXP3*, not included in  $\mathbf{Y}$ ) which showed decrease in expression along  $\mathbf{t}$  ( $|\rho_{\text{Spearman}}| = 0.56, 0.62, 0.59$ ) implying an association between the inferred trajectory and prognostic potential.

## 4 Discussion

We have proposed the concept of replacing discrete disease staging with data-driven continuous trajectories inferred from genomics data that hold prognostic and/or diagnostic value. By considering a modified factor analysis model we incorporate population-level heterogeneity that may confound existing trajectory-based analysis. By applying our model to lipid metabolic RNA-seq data from the TCGA colorectal cancer dataset we simultaneously identify genes that possibly interact with externally measured covariates while learning a trajectory that correlates with known markers of colorectal cancer prognosis.

One limitation of the model is its linear nature, making inferred latent variables similar to those from factor analysis. We therefore propose a nonlinear, nonparametric extension similar to Gaussian Process Latent Variable Models [11]. The trick is to consider the  $\mathbf{X}\mathbf{X}^T$  term in the covariance matrix of the marginal distribution of  $\mathbf{Y}$  and replace it with any (semi-)positive definite matrix representing “similarity” between the elements of  $\mathbf{y}$ , such as double-exponential kernels. We therefore mention the possibility of *Covariate Gaussian Process Latent Variable Models* with kernels given by

$$K(\{\mathbf{x}_{p=1,\dots,P}, \mathbf{t}\}, \{\mathbf{x}'_{p=1,\dots,P}, \mathbf{t}'\}) \propto K(\mathbf{x}, \mathbf{x}') + K(\mathbf{t}, \mathbf{t}') + \sum_p K(\mathbf{x}_p \odot \mathbf{t}, \mathbf{x}'_p \odot \mathbf{t}') \quad (6)$$

for some suitable choice of kernel function  $K$ . Note that the element-wise product  $\odot$  only appears because of the linear relationship between the covariates and factor loading matrix. This could easily be replaced by any nonlinear and possibly nonparametric function  $\mathbf{f}$  giving terms of the form  $K(\mathbf{f}(\mathbf{x}_p, \mathbf{t}), \mathbf{f}(\mathbf{x}'_p, \mathbf{t}'))$ .

A further limitation of the model is the current reliance on Gibbs sampling for inference, which scales poorly for larger numbers of samples and covariates. However, the complete conditional conjugacy of exponential family distributions make this model amenable to fast approximate inference methods for large datasets such as using Stochastic Variational Inference [12].

## References

1. Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature biotechnology* **32**, 381–386 (2014).

2. Reid, J. E. & Wernisch, L. Pseudotime estimation: deconfounding single cell time series. *Bioinformatics* **32**, 2973–2980 (2016).
3. Tatlow, P. & Piccolo, S. R. A cloud-based workflow to quantify transcript-expression levels in public cancer compendia. *bioRxiv*, 063552 (2016).
4. Weinstein, J. N. *et al.* The cancer genome atlas pan-cancer analysis project. *Nature genetics* **45**, 1113–1120 (2013).
5. Jones, R. G. & Thompson, C. B. Tumor suppressors and cell metabolism: a recipe for cancer growth. *Genes & development* **23**, 537–548 (2009).
6. Boland, C. R. & Goel, A. Microsatellite instability in colorectal cancer. *Gastroenterology* **138**, 2073–2087 (2010).
7. Lee, J.-S. & Surh, Y.-J. Nrf2 as a novel molecular target for chemoprevention. *Cancer letters* **224**, 171–184 (2005).
8. Khorrami, S., Hosseini, A. Z., Mowla, S. J. & Malekzadeh, R. Verification of ALDH activity as a biomarker in colon cancer stem cells-derived HT-29 cell line. *Iranian journal of cancer prevention* **8** (2015).
9. Park, J.-I. & Kwak, J.-Y. The role of peroxisome proliferator-activated receptors in colorectal cancer. *PPAR research* **2012** (2012).
10. Shang, B., Liu, Y., Jiang, S.-j. & Liu, Y. Prognostic value of tumor-infiltrating FoxP3+ regulatory T cells in cancers: a systematic review and meta-analysis. *Scientific reports* **5** (2015).
11. Lawrence, N. D. Gaussian process latent variable models for visualisation of high dimensional data. *Advances in neural information processing systems* **16**, 329–336 (2004).
12. Hoffman, M. D., Blei, D. M., Wang, C. & Paisley, J. W. Stochastic variational inference. *Journal of Machine Learning Research* **14**, 1303–1347 (2013).